Finishing *Drosophila mohavensis* Fosmid Clone 450-M05

Bill Dirkes

4/3/2008

Abstract

One of the goals of Bio 4342, Introduction to Genomics, is to finish and annotate the dot chromosomes of various species of fruit flies. Analysis of this data will hopefully provide insight into the genomic features involved in heterochromatin formation, including repetitious elements, genes, and non-coding RNAs. Currently Bio 4342 students are finishing the *Drosophila mohavensis* dot chromosome. In this paper I present my work on finishing project 450-M05.

Workflow

Initial Project Assessment

Assembly view of project 450-M05 initially showed four contigs, with discrepant forward-reverse pairs between contigs 3 and 5. Crossmatch showed sequence matches with contig 3 and all other contigs, as well as contig 5 with contigs 4 and 6c. The ends of the contigs as well as the left end of contig 3 had low read coverage.

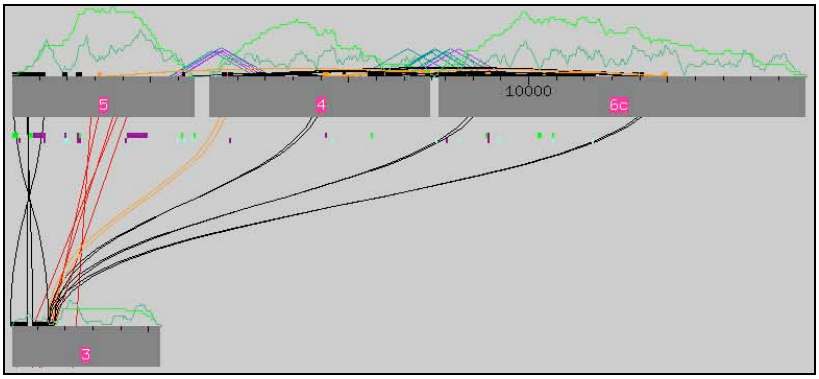

Figure 1: Initial Assembly View with Crossmatch results.

Initial Project Work

At first, I ignored contigs 3 and 5, and attempted to close the gap between contigs 4 and 6c. *Search for String* did not reveal any potential force join points, thus I abandoned this approach. Then, based on the Crossmatch results, it appeared that the left end of contig 3 matched with the left end of contig 5. However, before trying to join these two contigs, I inspected the discrepant forward-reverse pairs. I found that these reads not only faced the wrong direction, but they also contained multiple high quality discrepancies compared to the consensus sequence. Examination of the traces confirmed these high quality discrepancies. Thus, the reads in each pair with the greatest number of high quality discrepancies were placed into their own contigs.

Following the removal of the discrepant reads, the left end of contig 3 and the left end of contig 5 were joined (Figures 2 and 3). Final restriction digest data (Figures 15 and 16) support this join, showing the *in silico* fragments in the region are consistent with the *in vitro* results.

Figure 2: The force join between contigs 3 and 5.

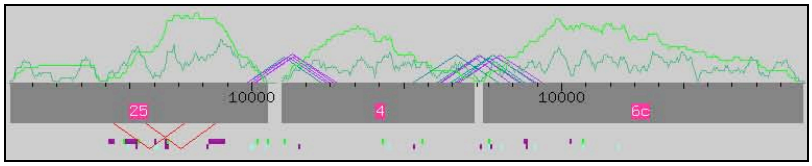

Figure 3: Assembly View after joining contigs 3 and 5.

I then attempted to join all reads in single-read contigs with the other contigs. However, the reads were reinserted into the same positions as before, as shown in Figure 4, and still contained high quality discrepancies. Thus, I removed these reads once again.

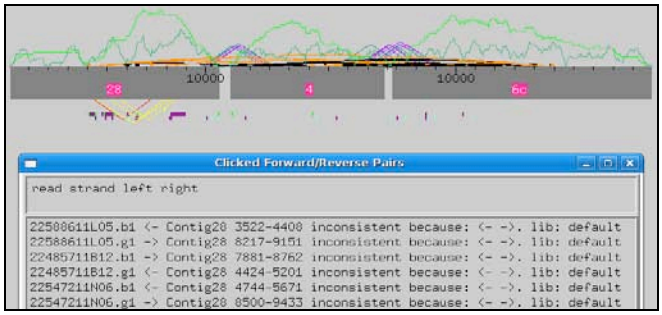

Figure 4: After reinsertion of reads in single-read contigs.

Following this, I improved the base-calling from trace data, both in regions near gaps and in low-quality regions on single-read contigs. Even after sequence improvement, I could not join any contigs.

Comparison between Autofinish and My 1st Round Primers

I chose multiple primers to confirm the accuracy of the force join and span the remaining gaps.

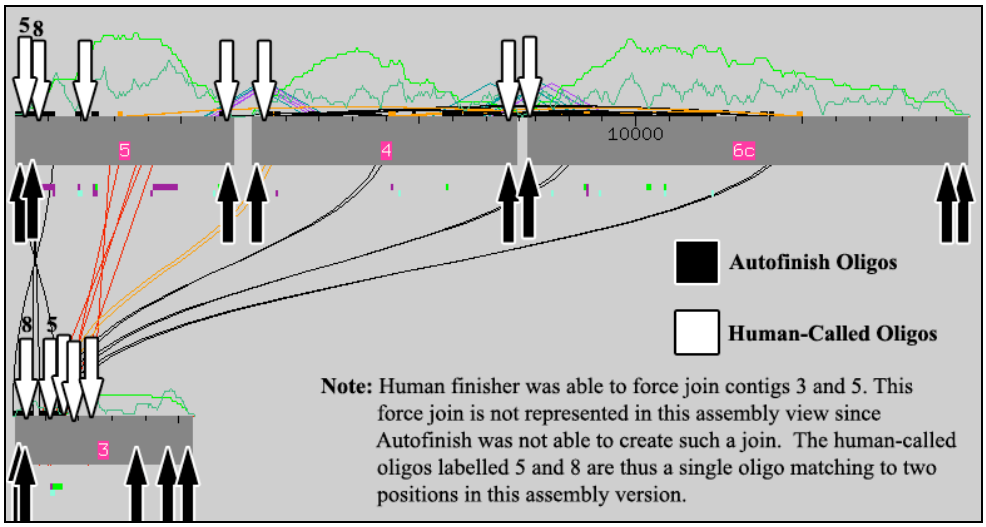

Figure 5: Comparison of oligonucleotide locations between Autofinish and student finisher.

Autofinish called thirteen primers, whereas I only called ten. While Autofinish called fewer reactions near the ends of contigs 3 and 5, the program did not appear to recognize the clone ends, and thus called multiple reactions at the ends of the project. While Autofinish and I did not have any identical oligonucleotides (Tables 1 and 2), the oligonucleotides picked were in very close proximity (Figure 5).

oligoID	oligoName	oligoSequence	Position (C=Contig)
2167	450-M05.1	gctgcctggtaagctaactgt	C4 7758-7778
2166	450-M05.10	gcatttaagcgcaccattac	C3 2429-2448
2165	450-M05.9	tgatagggaagattcttttgg	C3 1950-1970
2171	450-M05.7	aatgccaaaagtcgtatccta	C6 13206-13226
2170	450-M05.3	gtttctgaacaatcaaaagtgttat	C5 6360-6384
2169	450-M05.6	gaacttttcaaagcgtttgtc	C5 2124-2144
2168	450-M05.2	ccatcttacatgggagtcttaaat	C4 291-314
2164	450-M05.8	gccaggatttcgatctgta	C3 390-408, C5 804-822
2163	450-M05.5	aaagcacaagtcaaatgtgag	C3 1075-1095, C5 306-326
2162	450-M05.4	cactctcattttattccaagaag	C3 1652-1674

Table 1: My oligonucleotides for the first round of finishing

(primer bases)	(melt temp)	(strand)	(first base of read)	(contig)
Ccttgctgttcccacga	59	<-	165	Contig3
Ggcggttcgctcagata	58	<-	366	Contig3
Attggctctagcagtaattatttta	55	<-	3770	Contig3
Agtaacaaacagcacgagtgta	55	->	4739	Contig3
Cgactagagcgttcttacttgttt	58	->	5331	Contig3
Cgctgtagtgtgggca	56	<-	91	Contig4
Ggcctcagcattaccaat	56	->	7985	Contig4
Gcgctgtaattacgaacatt	56	<-	95	Contig5
Accagcgaaccagcata	56	<-	685	Contig5
Gctgaaagagcgattaccc	57	->	6442	Contig5
Aagttcagccttttacttgatg	55	<-	184	Contig6
Tgaaatttaccagttttgatagaaa	56	<-	774	Contig6
Ggatacgggtctgttatcg	55	->	13315	Contig6

Table 2: Autofinish's oligonucleotides for the first round of finishing

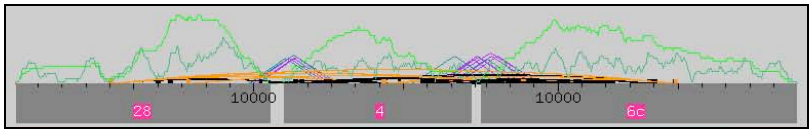

Figure 6: Assembly View after 1st round reads added.

After First Round of Additional Sequencing Reactions

The Assembly View, following addition of the new first round reads (Figure 6), revealed little change in the scaffold. No gaps could be closed based on string searches. However, careful examination of the reads revealed that two of the new reads were misassembled. The reads had been inserted near the center of contig 6c, whereas they belonged at the right end of contig 4. Pulling out the reads and inserting them into the proper position extended contig 4 far enough to allow me to join the two contigs.

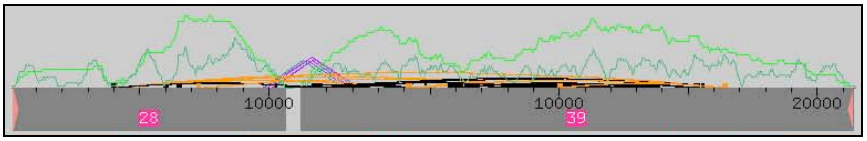

Figure 7: Assembly View after the force join.

After making the join (Figure 7), I continued to evaluate high quality discrepancies, marking putative polymorphisms. In addition, I identified regions of low quality and designed primers for sequencing reactions to increase the Phred score in these regions.

After 2nd Round of Additional Sequencing Reactions

Although adding the new reads did not appear to change the assembly view, one read did extend further into the remaining gap, and the consensus Phred score near the original force join within contig 28 increased dramatically.

In addition, two new reads contained multiple high quality discrepancies. These reads were removed, however subsequent attempts to rejoin the reads elsewhere failed. Another read, 0373*L15.g1, also contained high quality discrepancies, and its forward-reverse pair was approximately 3kb from the left end of contig 39, facing the gap. This led to the hypothesis that the read belonged within the gap. However, the *Search for String* function did not find any overlap between the edges near the gap. Thus, I designed two primers to read off the ends of this read, in addition to multiple primers at the ends of contigs 28 and 39 facing the gap.

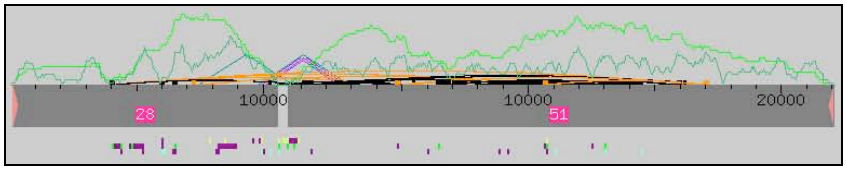

Figure 8: Assembly View before adding new reads. Note that contig 51 has been extended by approximately 700bp by moving misassembled reads.

After 3rd Round of Additional Sequencing Reactions

The final round of finishing included primers to cover low quality regions, as well as multiple primers to span the gap, as previously described. Although fifteen new reads were added, Assembly View appeared the same as before. In addition, the reactions off 0373*L15.g1 failed, and thus we could not use the read to create a force join. Comparing read name to the primers revealed that many reads were misplaced. However, *Search for String* did not reveal any valuable rearrangements of the reads.

Because no further reactions could be called and one gap remained, the consensus sequence of project 455_J01 was used as a reference. 455_J01 is estimated to overlap all but 11kb of project 450_M05. However, 455_J01 is believed to contain a 15kb polymorphism, and thus may not be an ideal reference sequence.

Figure 9: Low quality join created. More data needed to improve Phred score.

Initially, I identified the region of 455_J01 that mapped to the gap. GSC finisher C. Strong created a low-quality join by using the reference sequence and trace data to improve base calls in the region (Figure 9). To gain confidence in the join, I attempted to add the consensus sequence from 455_J01 as a fake read. I broke the original consensus into four similarly-sized pieces with approximately 60 overlapping bases. This reduces the memory overhead required by PhredPhrap. However, this approach did not work, as shown in Figure 10, because rerunning PhredPhrap caused the misassemblies solved earlier to resurface.

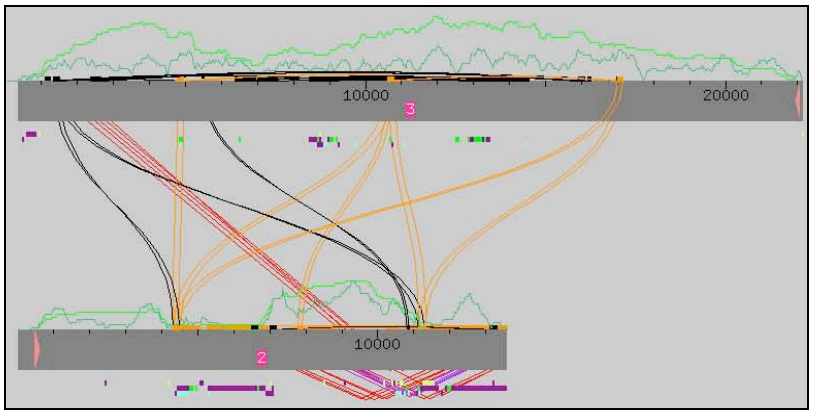

Figure 10: Assembly View after adding consensus from 455-J01 and rerunning phredPhrap.

As a second attempt, I created a 400bp fake read titled 455_J01.c1 centered on the gap. This read was added to the version of the project immediately prior to the addition of the previous fake reads. Consed added the fake read to a different region than the intended gap. Thus, the fake read was removed and rejoined into the proper location (Figure 11).

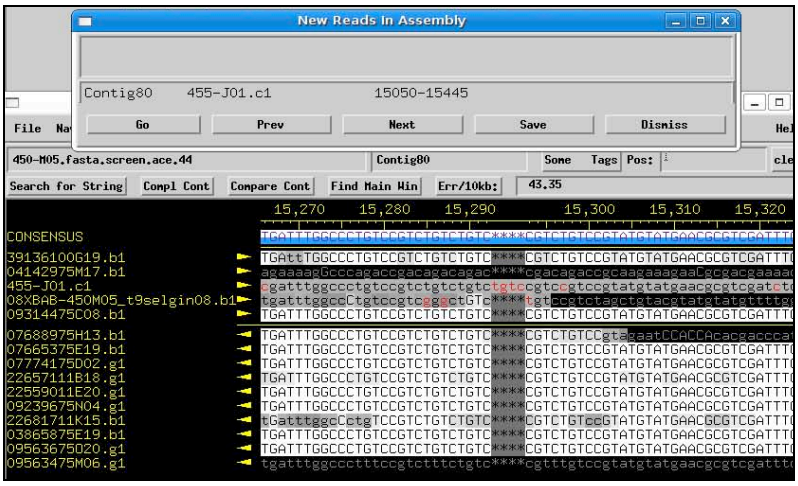

Figure 11: Region where "Add New Reads" placed the fake read 455-J01.c1

Due to the low quality of the consensus, even after the join, further reactions are needed to verify the sequence. Furthermore, the join contained many discrepancies, probably due to the very low quality in the consensus sequence (Figure 12).

Figure 12: Join between fake read and consensus.

Final Work

Finally, I removed three 3rd round reads due to multiple high quality discrepancies at different positions, as shown in Figure 13. All were different reaction chemistries using the same oligonucleotide primer, identifiable by 08X*13selgin08.b1 in the *Find Reads* function of Consed. Trace data showed two overlapping, high quality bases at multiple positions, suggesting that the oligonucleotide primed at two locations (Figure 13). The *Search for String* function was used prior to the sequencing reactions, and thus this probably resulted because the final ten nucleotides of the oligonucleotide sequence matched to another region in the project. This can result in multiple priming locations, even if the entire oligonucleotide sequence is unique.

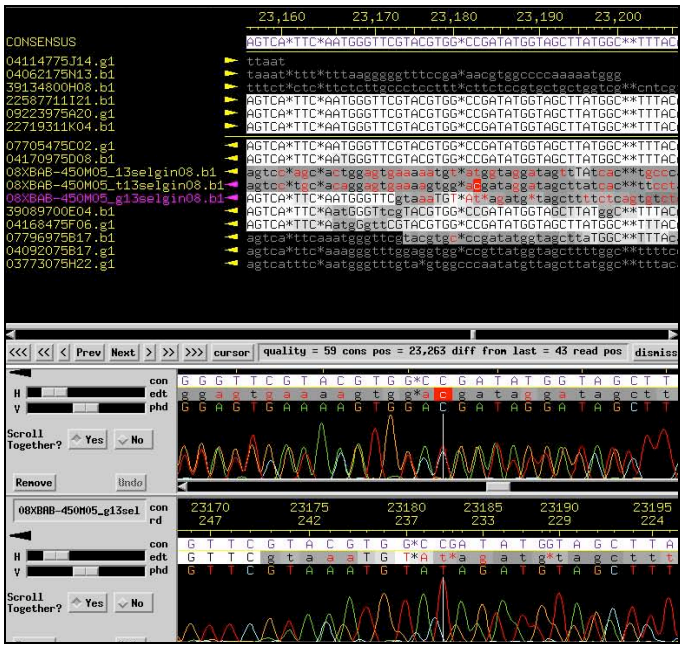

Figure 13: Trace data suggests multiple priming loci.

In addition to high quality discrepancies, I tagged three putative polymorphisms. This conclusion was based on examination of all read traces, both at the immediate base position and surrounding regions. All contained high quality traces with no other discrepancies nearby. In addition, one mononucleotide run was identified and tagged. The traces of both the mononucleotide run and the flanking regions were checked, with no observable miscalled bases.

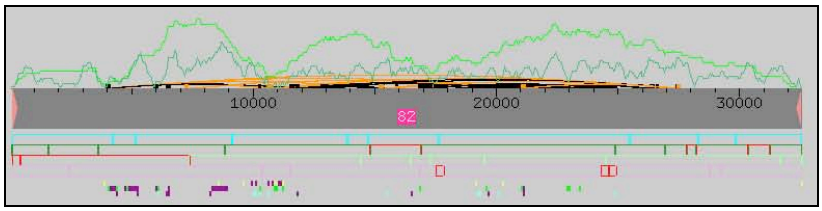

Figure 14: Final Assembly View. Digest data supports the final consensus sequence.

Analysis of Digest Data

The digest data from the project provides additional supporting evidence for the final sequence consensus (Figures 14 and 15). *EcoR*I and *Sac*I digests match with *in silico* predictions. In addition, the *in vitro Hind*III digest matches with *in silico* predictions except for a single missing *in silico* band around 7kb. Two digests agree with the *in silico* results, with two other digests having very similar banding patterns. Thus, the digest data supports the consensus sequence.

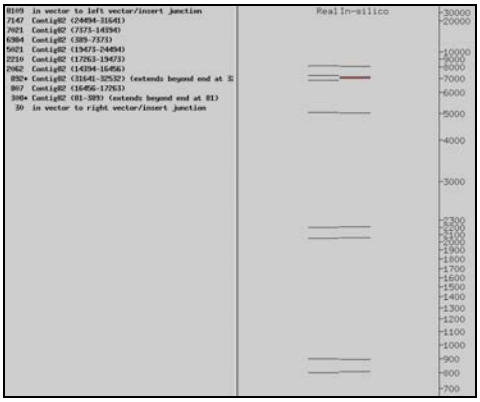

Figure 15: *EcoR*I digest. See Figure 17 for evidence that the red (discrepant) in-silico fragment is actually a doublet.

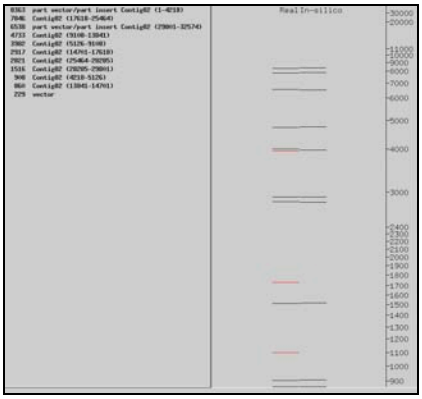

Figure 16: *EcoR*V digest. Two *in vitro* fragments do not have a corresponding *in silico* fragment. Because other digests show far better matches (see Fig. 15), this digest alone is not indicative of a misassembly, but suggests that a base in the restriction site was miscalled during finishing.

Further Work

Cloning ends have been tagged, a single large primary contig remains, and BLAST was run to verify no contamination occurred. As described previously, three putative polymorphisms and a mononucleotide run were tagged.

However, the project requires further work as it does not meet all specifications yet. Two low quality regions remain, both near the location of the last gap closed. Due to the repetitious sequence in this region, alternative techniques should be utilized (such as PCR to generate a specific template, rather than using the fosmid template). All other regions meet the minimum specification of Phred 25 for double-stranded and Phred 30 for single-stranded regions. It should also be noted that the consensus sequence from project 455-J01 contains some discrepancies from the consensus of this project, which may be reason for concern. However, the digest data increases my confidence in this consensus sequence.

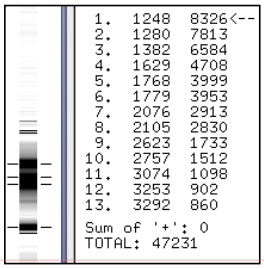

Figure 17: *EcoR*I digest image. The resolution is not high enough for the computer to identify the doublet at approximately 7000bp. See Fig. 15 for *in silico* predictions.

SCR Elgin 4/20/08 5:53 PM
Deleted: c

Special Thanks

Special thanks to GSC finisher C. Strong, who assisted with a significant portion of the project. In particular, her expertise made the final join possible. In addition, I would like to thank GSC finisher L. Courtney, who assisted in the initial stages of fixing the misassembly between contigs 3 and 5.

Supplemental Data

Project files, screenshots, consensus sequence of 455-J01, fake read sequence, and digest images are available upon request.